ZHIYI YUE

Phone: 2812361357 | Email: Zhiyi.Yue@uth.tmc.edu

Education

The University of Texas Health Science Center at Houston *09/2023-Present*
Doctor of Philosophy in Biostatistics
Core Courses: Statistical Inference, Linear Models, Survival Analysis, Stochastic Process in Biostatistics, Health Economics

Columbia University, New York, NY *09/2021-02/2023*
Master of Art in Statistics
Core Courses: Probability, Linear Regression Models, Computational Statistics and Introduction to Data Science, Computer Systems for Data Science, Advanced machine learning, Game theory

Beijing Normal University - Hong Kong Baptist University United International College(UIC) *09/2017-06/2021*
Bachelor of Science (Honours) in Statistics
Honors: UIC First-Class Scholarship (11/2020); UIC Second-Class Scholarship (11/2019 & 11/2018)

Research Interests

biostatistics, disease detection, epidemiological methods, clinical trials, health equity, survival analysis, data mining, machine Learning

Core Competencies & Skills

- **Programming languages & Tools**: R, Matlab, Python, C, SPSS, SQL, QGIS, Microsoft Office (Word, PowerPoint, Excel, Access)
- **Methodologies:** Machine Learning, Cross Validation, Predictive Modeling, Statistical Analysis, Feature Engineering, Causal inference, Hypothesis Testing, A/B Testing, Data Management, Data Visualization
- **Languages**: Mandarin (Native); English (Fluent): TOFEL 101, GRE 324

Research Experiences

Research Project - The Impact of Fracking Wells on Adverse Neonatal Outcomes *03/2024-Present*

- Conducted a literature review to identify pollutants and toxic substances released by fracking wells, highlighting their potential links to diseases often overlooked in existing research, thereby identifying research gaps.
- Developed a research framework to investigate the racial and ethnic disparities in adverse birth outcomes, emphasizing the need for equitable health protections for vulnerable communities.

Research Project - Data Analysis on Racial Disparities in Urban Trail Use *10/2023-03/2024*

- Utilized R to build logistic regression models assessing sociodemographic characteristics related to urban trail use, adjusting for variables such as location, day of the week, and time of day.
- Performed statistical analysis to determine the likelihood of different racial groups participating in various activities (e.g., walking, running, cycling) on the trail, interpreting odds ratios and p-values to identify significant disparities.
- Analyzed data to reveal significant racial and ethnic disparities in trail use, providing insights into how different groups engage with urban physical activity infrastructure.

Research Project - Functional Data Analysis based on Septic System Locations and Conditions *06/2022-12/2022*

- Collected geographic data like depth to groundwater, the locations of buildings in Georgia and census data through USGS and US census
- Analyzed geographic data by using QGIS for statistical model building and conducted exploratory data analysis to find the relationship between these variables
- Conducted data analysis model like logistic regression and random forest to predict whether septic systems need repairment and explore determinants of septic failure

Research Project - Detection of Diabetes Based on Generative Adversarial Network *07/2022-09/2022*

- Analyzed diabetes records dataset with 5070 samples and 9 covariates including sex, year of birth, body mass index, family history of diabetes, diastolic blood pressure, oral glucose tolerance test, insulin releasing test, triceps skinfold thickness and diabetes class
- Implemented traditional classification techniques like K-nearest neighbor and Decision tree by Python to make predicted classification
- Applied an improved model DiGAN, which combines Generative adversarial network and Random forest, can achieve an F1-score up to 98%

- Used Local interpretable model-agnostic explanations and SHapley additive explanations to interpret the model and determine the most important features

Undergraduate Thesis - Covid-19 Transmission Risk Predictors *06/2020-12/2020*

- Collected climate and air pollutant data of 15 cities and use 8 classification methods in R including Random Forest and SVM to classify these cities according to the degree of COVID-19 transmission risk
- Used 4 neural network algorithms in Python like RNN and LSTM to make predictions of transmission trend of disease, including optimization of hyperparameters such as batch size
- Utilized feature importance algorithms such as Gini index to determine meteorological and epidemiological factors most responsible for COVID-19 transmission risk
- Minimized overfitting through traditional machine learning data preprocessing methods such as standardization, boosting, and bagging

Research Project - Google App Store Business Analysis *07/2019-09/2019*

- Constructed multivariate linear regression models and random forest models to investigate factors responsible for positive customer reviews for apps in the Google Play Store
- Prepared developmental proposals through trend analysis to inform app developers on the characteristics behind popular apps
- Created data visualizations such as pair plots and histograms for exploratory data analysis in Python and Matlab to extract customer behavior patterns in Google App Store

Publications

- Zhao, P., Liu, X., Yue, Z., Zhao, Q., Liu, X., Deng, Y., & Wu, J. (2024). DiGAN Breakthrough: Advancing diabetic data analysis with innovative GAN-based imbalance correction techniques. *Computer Methods and Programs in Biomedicine Update*.

Oral Presentations

- **Yue, Z.**, Chen, B., Cofer, H., Jensen-Morgan, A., McMichael, C., Salinas, J., Vargas, S., Young, C., & Lanza, K. (Accepted). Racial and ethnic disparities in urban trail use in Central Texas. American Public Health Association Annual Meeting. Minneapolis, MN.
- Lanza, K., **Yue, Z.**, & Chen, B. Trail sociodemographic inventory study. (2024). The Trail Conservancy 2024 Board Retreat. Lady Bird Johnson Wildflower Center. Austin, Texas.

Professional Experiences

Deloitte Consulting *07/2020 - 08/2020*

Part-Time Assistant - Consulting Analyst Intern

- Developed diverse data visualizations through PyChart such as bar charts and trend analyses for presentation to senior management to inform future sales decision making and marketing strategies
- Conducted marketing analysis and market research to understand trends in the supplement healthcare industry
- Extracted data to determine seasonality effects and sales trends through data analysis in R
- Investigated factors behind sales performances and distribution channel cost effectiveness in terms of market scale and future buying power trends

Roche - China *12/2019 - 01/2020*

Sales Analyst Intern

- Developed sales reports and budget outlines in SQL and Excel for financial planning and reporting
- Gathered data on medical drugs and equipment expenditures and sales volumes to inform pricing strategies and market knowledge
- Employed statistical methods to analyze trends in sales performance and other KPIs to pinpoint the optimal marketing portfolios among different product lines

Extracurricular Activities

Gapper International Volunteer Organization *06/2019*

Volunteer

- Taught primary school students English and mathematics in Langkawi, Malaysia

Dream Safari Charitable Organization of UIC *09/2018-06/2019*

Minister of HR Department

- Held department meetings regularly and dealt with other administration works
- Took charge of the recruitment of the organization and organized various conferences

Shanghai Adream Foundation Dream Coach Program *07/2018*

Volunteer

- Trained front-line teachers at the local schools of two cities in Yunnan China, delivered cutting-edge education concept and presented teaching methods of the courses